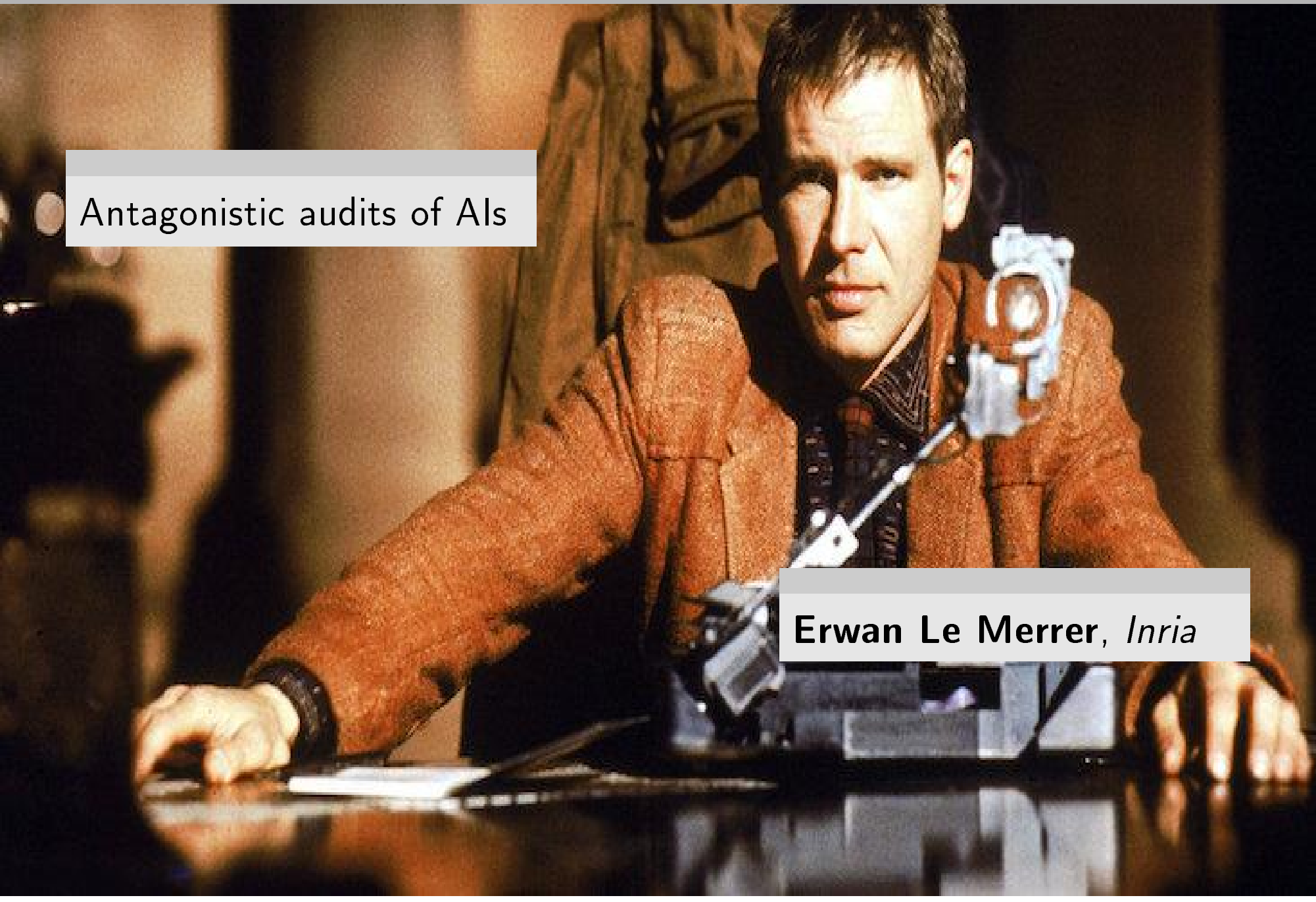


Antagonistic audits of AIs

Erwan Le Merrer, *Inria*

Today: ChatGPT or student?

Sung Kim

Dec 11, 2022 · 4 min read · Member-only · Listen

How to Detect OpenAI's ChatGPT Output

How to detect if the student used OpenAI's ChatGPT to complete an assignment

On November 30, 2022, OpenAI released 'ChatGPT' AI system (https://openai.com/blog/chatgpt/), which is a universal writer's assistant that can generate a variety of output, including school assignments. The output (e.g., essays) provided by ChatGPT is so good, if I was a student, I would be using ChatGPT to complete most of my school assignment with minor revisions.

Today: ChatGPT or student?

Can AI-Generated Text be Reliably Detected?

Vinu Sankar Sadasivan
vinu@umd.edu

Aounon Kumar
aounon@umd.edu

Sriram Balasubramanian
sriramb@umd.edu

Wenxiao Wang
wwx@umd.edu

Soheil Feizi
sfeizi@umd.edu

Department of Computer Science
University of Maryland

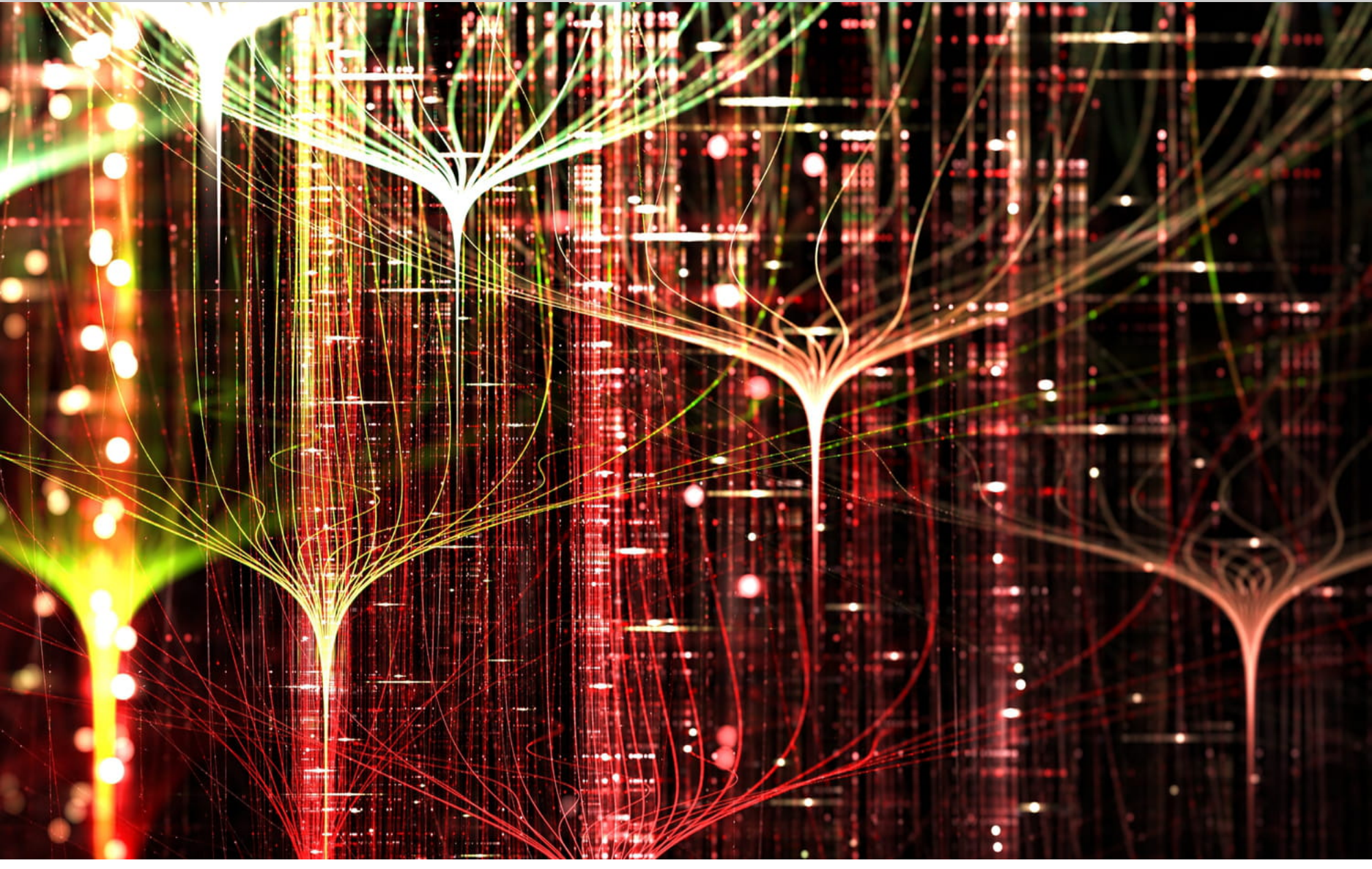
Abstract

The rapid progress of Large Language Models (LLMs) has made them capable of performing astonishingly well on various tasks including document completion and question answering. The unregulated use of these models, however, can potentially lead to malicious consequences such as plagiarism, generating fake news, spamming, etc. Therefore, reliable detection of AI-generated text can be critical to ensure the responsible use of LLMs. Recent works attempt to tackle this problem either using certain model signatures present in the generated text outputs or by applying watermarking techniques that imprint specific patterns onto them. In this paper, both empirically and theoretically, we show that these detectors are not reliable in practical scenarios. Empirically, we show that *paraphrasing attacks*, where a light paraphraser is applied on top of the generative text model, can break a whole range of detectors, including the ones using the watermarking schemes as well as neural network-based detectors and zero-shot classifiers. We then provide a theoretical *impossibility result* indicating that for a sufficiently good language model, even the best-possible detector can only perform marginally better than a random classifier. Finally, we show that even LLMs protected by watermarking

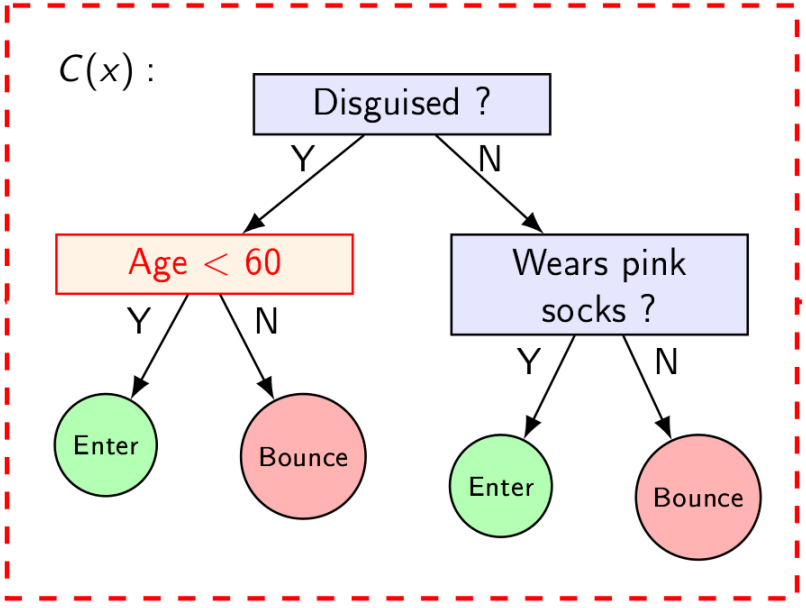
New Inria/IRISA team in Rennes

Joint research w. Gilles Tredan (CNRS)
Ph.D. students: Augustin, Gurvan, Timothée and Jade

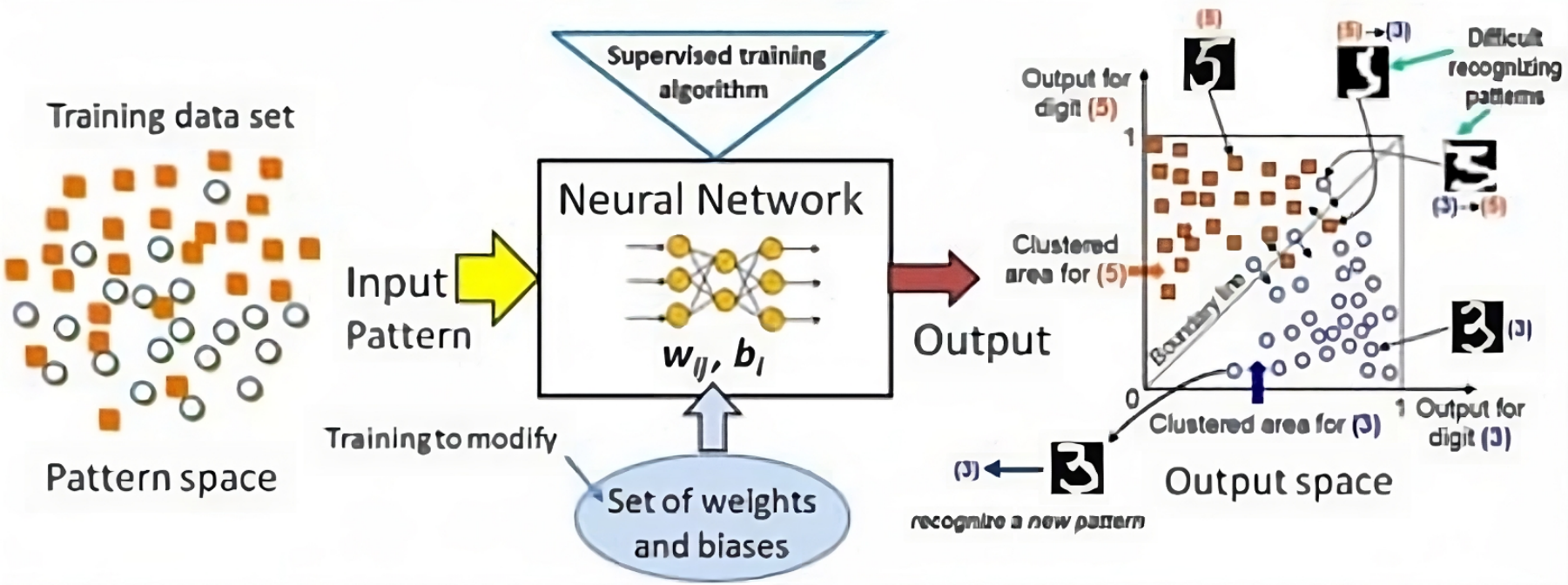
Algorithms solve tasks



Before AI: natively explainable algorithms

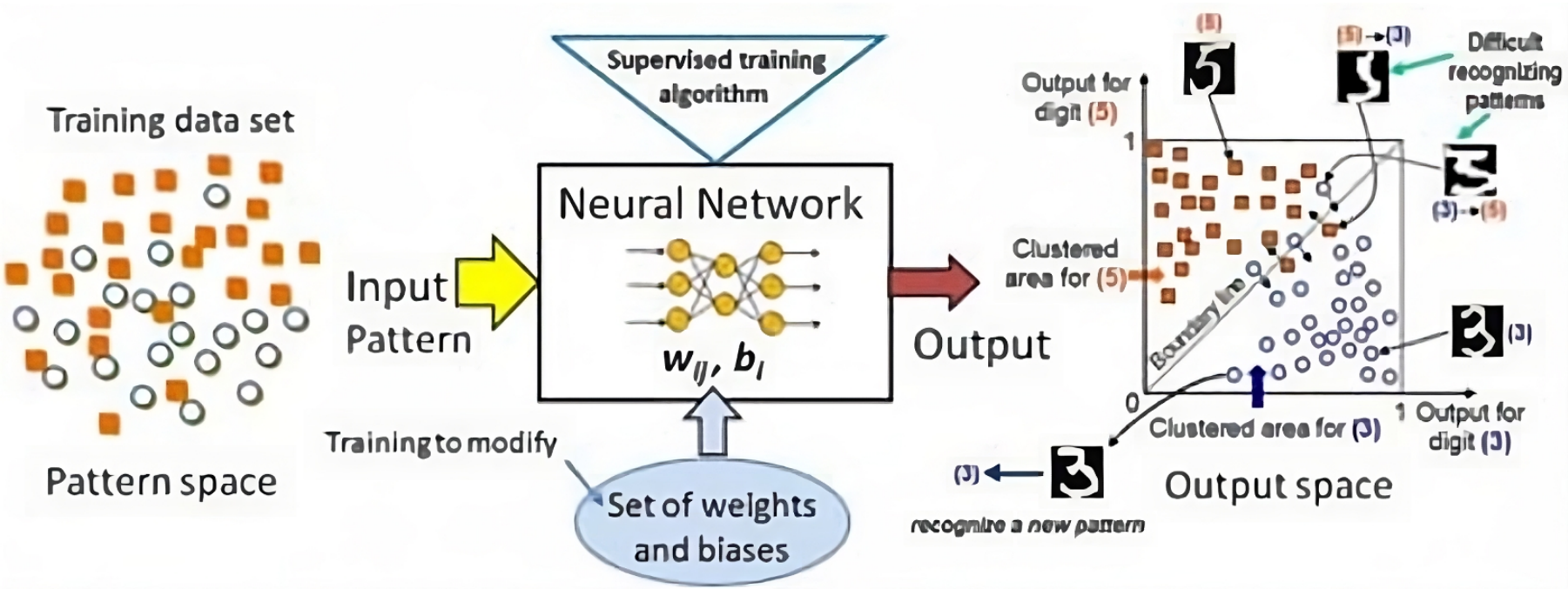


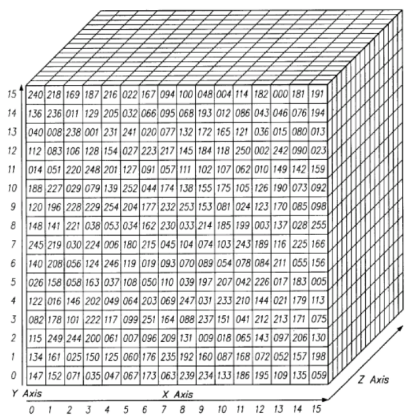
With AI: black-boxes (classification)



img: Le Dung et al. 2008.

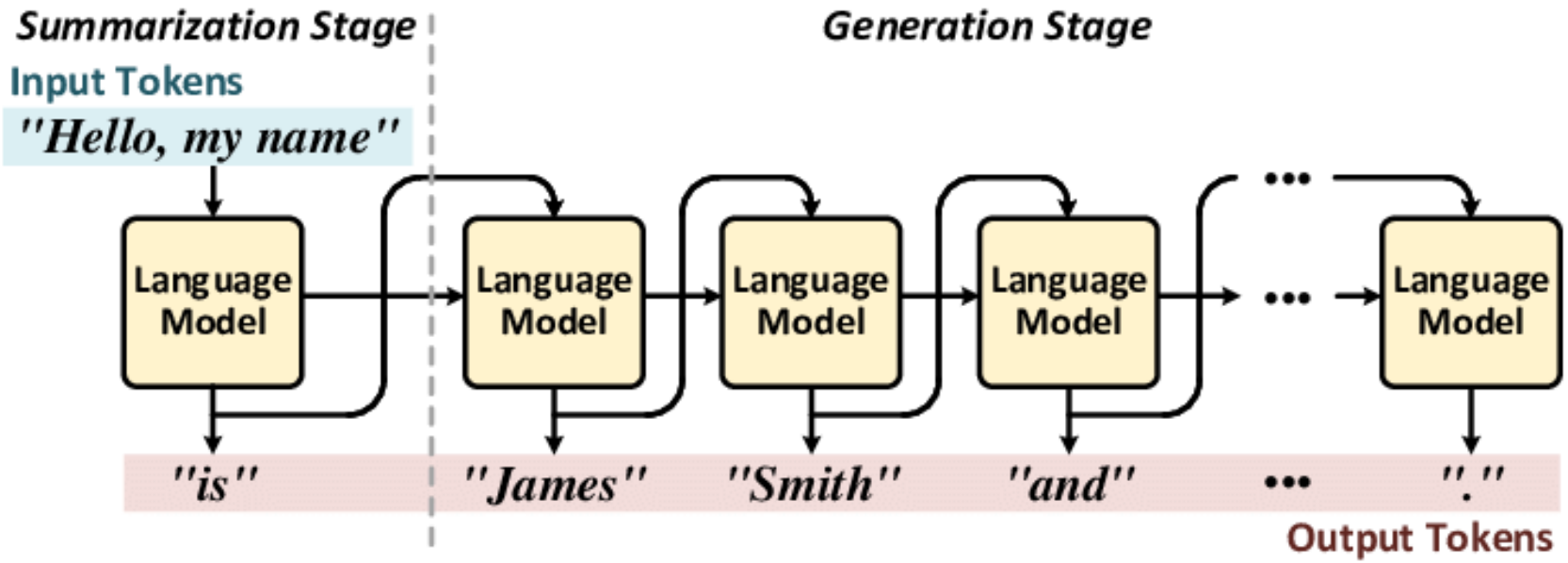
With AI: black-boxes (classification)





img: Le Dung et al. 2008.

Generative AIs



Multiple biases

LE MONDE *diplomatique*

NUMÉRO DU MOIS ARCHIVES CARTES AUDIO MANIÈRE DE VOIR HORS-SÉRIES BLOGS À PROPOS

> **Novembre 2024**, pages 8 et 9, en kiosques

1 traduction

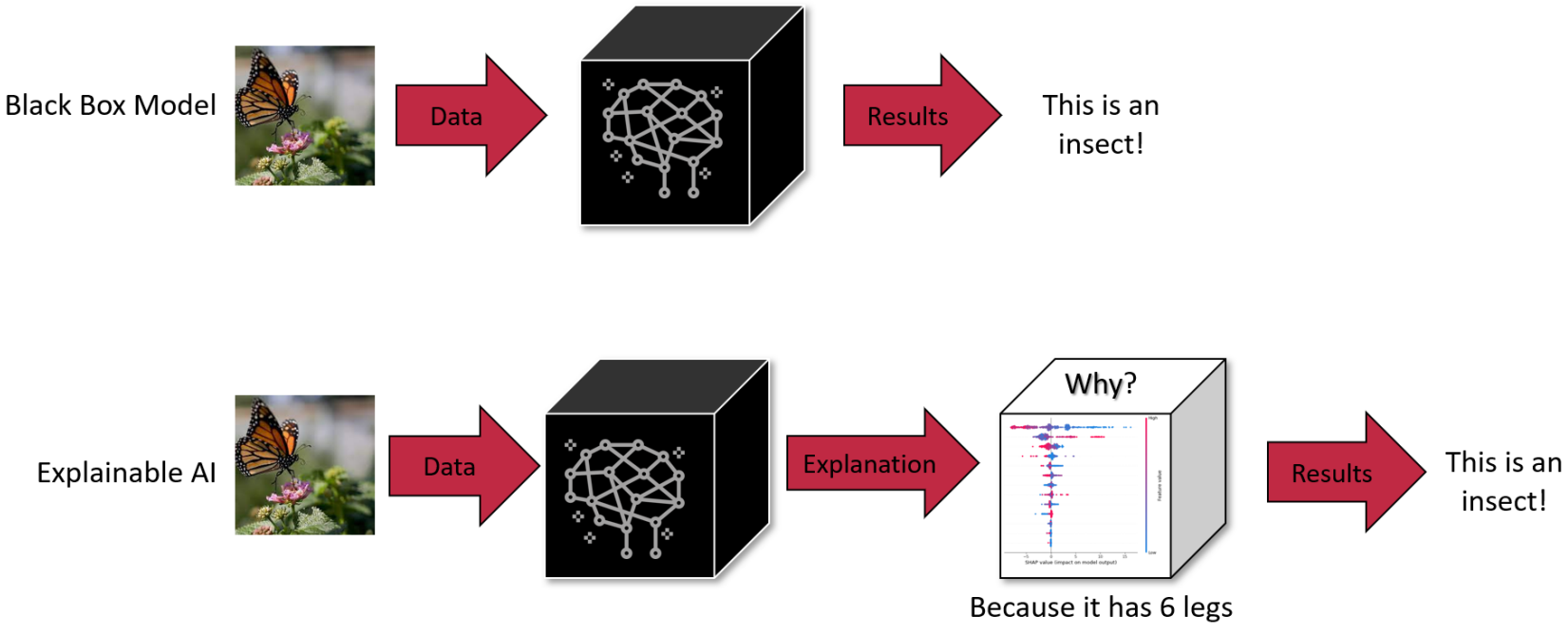
LA TECHNIQUE, C'EST TOUJOURS POLITIQUE

Pourquoi l'intelligence artificielle voit Barack Obama blanc

Quoi de plus neutre, dit-on, qu'un ordinateur ? Erreur : derrière leurs verdicts froids, algorithmes et automates encapsulent tous les biais des humains qui les conçoivent. Basée sur le modèle de l'individu calculateur, héritière d'une histoire tissée de choix idéologiques, l'intelligence artificielle est une machine politique. La mettre au service du bien commun implique d'abord de la déconstruire.

As an AI developer: dig to explain (XAI)

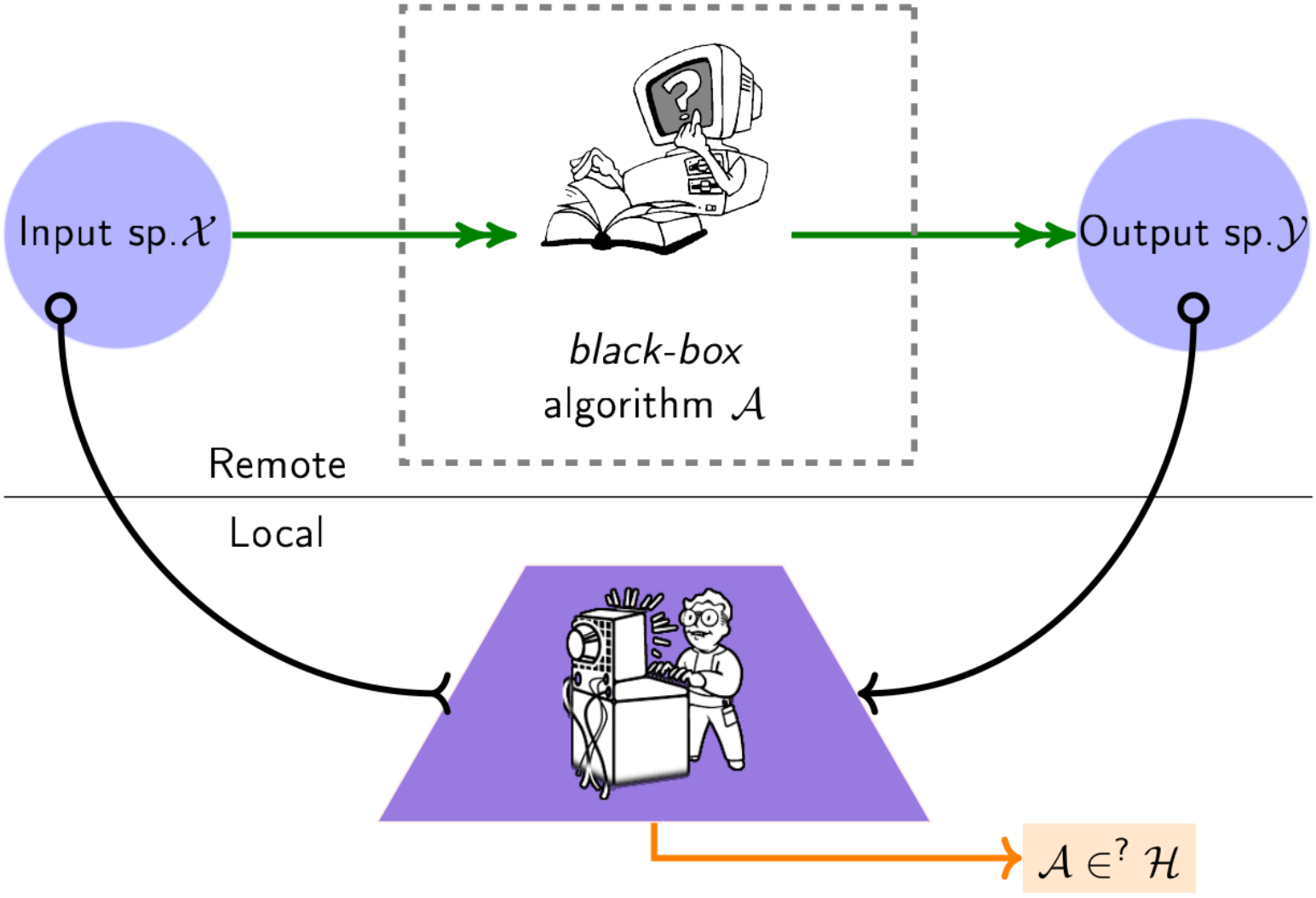
If "physical" access to the AI model:



img: Lundberg SM et al., 2017

Blade Runner: the Voight-Kampff test

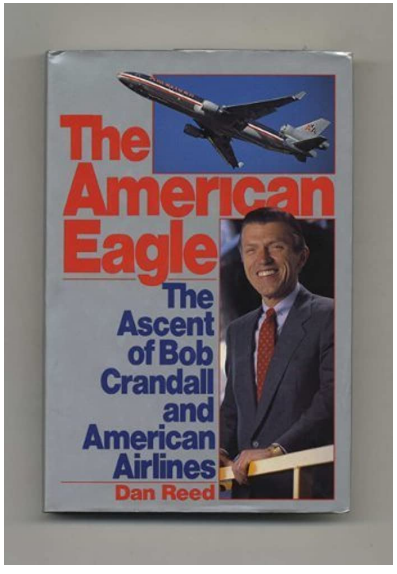
Is the **remote** entity a replicant ?
Essentially: investigation w. questions/answers (inputs/outputs)



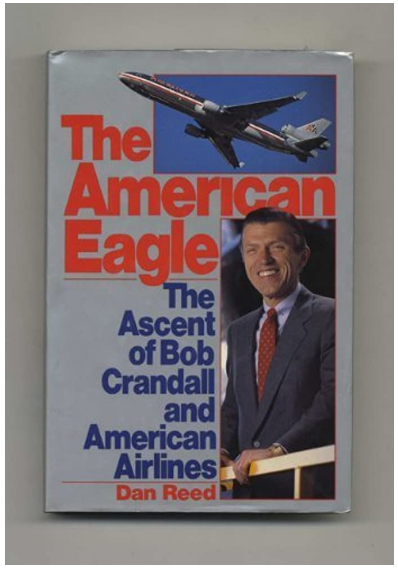
If $\mathcal{A}$ is truthful, then $\equiv$ XAI. But...

In summary: some black-box audits might work, provided the platform collaborates

In a pre-"Public Relations" world



In a pre-"Public Relations" world



- *Crandall's complaint* (A. airlines pres.) at congress (1983): **"Why would you build and operate an expensive algorithm if you can't bias it in your favor?"**

US Civil Aeronautics Board: 'screen bias' made illegal (1984)

The **Volkswagen emissions scandal**, sometimes known as **Dieselgate**[24][25] or **Emissionsgate**,[26][25] began in September 2015, when the United States Environmental Protection Agency (EPA) issued a notice of violation of the Clean Air Act to German automaker Volkswagen Group.[27] The agency had found that Volkswagen had intentionally programmed turbocharged direct injection (TDI) diesel engines to activate their emissions controls only during laboratory emissions testing, which caused the vehicles' NO_x output to meet US standards during regulatory testing. However, the vehicles emitted up to 40 times more NO_x in real-world driving.[28] Volkswagen deployed this software in about 11 million cars worldwide, including 500,000 in the United States, in model years 2009 through 2015.[29][30][31][32]

Volkswagen emissions scandal

A 2010 Volkswagen Golf TDI displaying "Clean Diesel" at the Detroit Auto Show

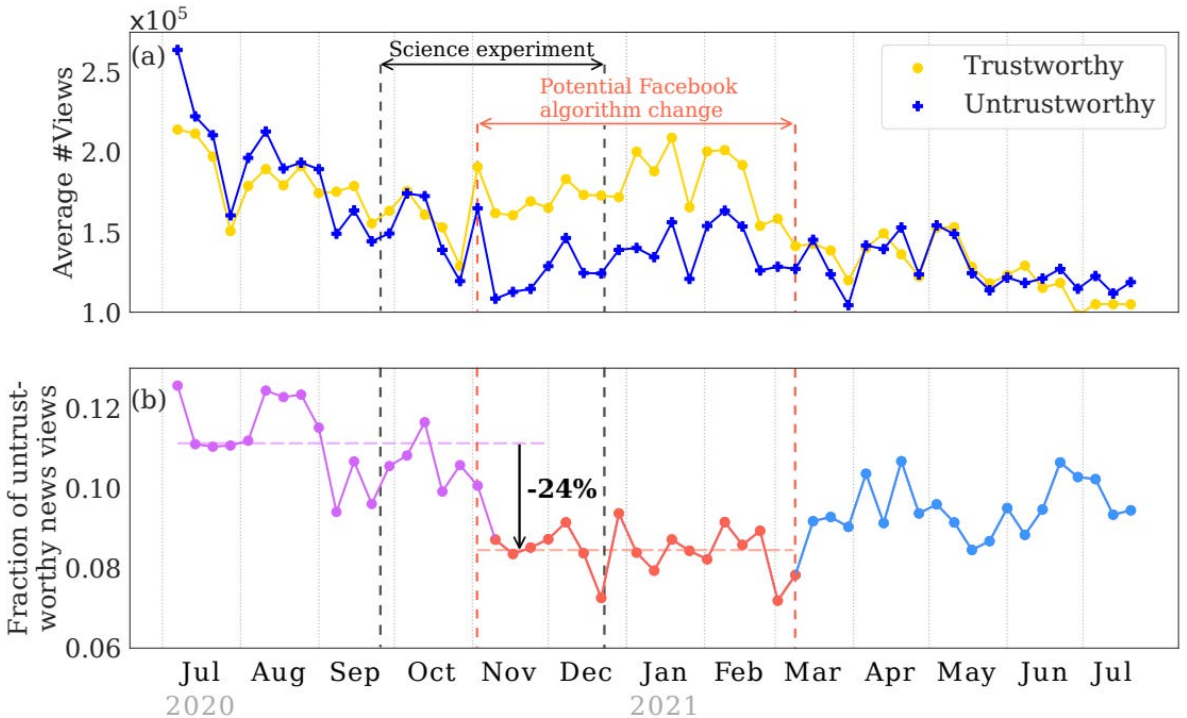
Date	2008–2015
Location	Worldwide
Also known	Dieselgate, Emissionsgate

The **Volkswagen emissions scandal**, sometimes known as **Dieselgate**[24][25] or **Emissionsgate**,[26][25] began in September 2015, when the United States Environmental Protection Agency (EPA) issued a notice of violation of the Clean Air Act to German automaker Volkswagen Group.[27] The agency had found that Volkswagen had intentionally programmed turbocharged direct injection (TDI) diesel engines to activate their emissions controls only during laboratory emissions testing, which caused the vehicles' NO_x output to meet US standards during regulatory testing. However, the vehicles emitted up to 40 times more NO_x in real-world driving.[28] Volkswagen deployed this software in about 11 million cars worldwide, including 500,000 in the United States, in model years 2009 through 2015.[29][30][31][32]

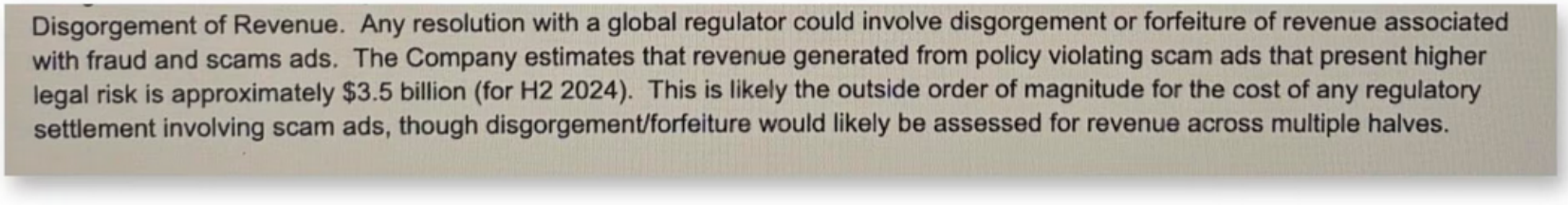
Volkswagen emissions scandal

A 2010 Volkswagen Golf TDI displaying "Clean Diesel" at the Detroit Auto Show

Date	2008–2015
Location	Worldwide
Also known	Dieselgate, Emissionsgate



How do social media feed algorithms affect attitudes and behavior in an election campaign? Science, 2023.


Disgorgement of Revenue. Any resolution with a global regulator could involve disgorgement or forfeiture of revenue associated with fraud and scams ads. The Company estimates that revenue generated from policy violating scam ads that present higher legal risk is approximately $3.5 billion (for H2 2024). This is likely the outside order of magnitude for the cost of any regulatory settlement involving scam ads, though disgorgement/forfeiture would likely be assessed for revenue across multiple halves.

An excerpt from a November 2024 strategy document discussing Meta's scam ad revenue and legal risks. Screenshot via REUTERS

Meta has also placed restrictions on how much revenue it is willing to lose from acting against suspect advertisers, the documents say.

https://www.reuters.com/investigations/meta-is-earning-fortune-deluge-fraudulent-ads-documents-show-2025-11-

In practice: in machine learning for production systems, utility is often clear, *e.g.*,:

- YouTube recommender system: maximizing per-user watch time (Recsys 2016)
- Facebook ads: accuracy of user-click prediction on candidate ads (ADKDD 2014)

i.e., not necessarily aligned with auditors metrics (fairness, diversity)...
$\implies$ satisfying auditors metrics might degrade platforms utility;
"**antagonism**: 2 metrics not optimized jointly & leading to utility degradation?"

Manipulation of the platform 1/2: faking fairness

Audit scheme:
`auditor` asks a `platform` a benchmark dataset Z to assess fairness (with *e.g.*, disparate impact metric)

Biased sampling attack:

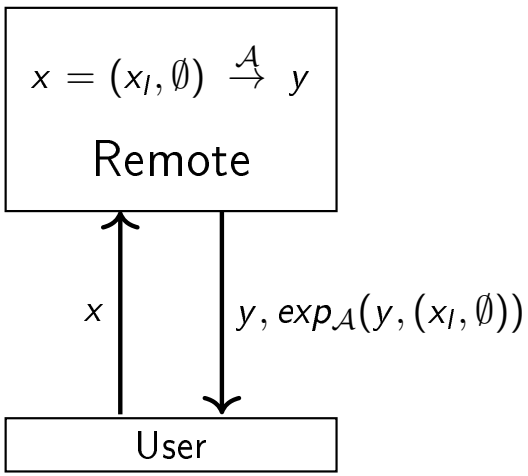
- `platform` has $D \sim P$ (underlying distribution, w. decisions)
- D may be unfair w.r.t. `auditor`'s metric
- `platform` selects $Z \subseteq D$ so that it is fair
- Z is given to `auditor`

`auditor` is manipulated (Z is indistinguishable from an originally fair dataset)
There exists and efficient algorithm for `platform` for sampling stealthily (reduction to min-cost flow problem)

Faking Fairness via Stealthily Biased Sampling, Fukuchi et al. AAAI 2020.

Manipulation of the platform 2/2: the bouncer problem

Audit scheme:
`auditor` finds discriminations in `platform` explanation of decisions

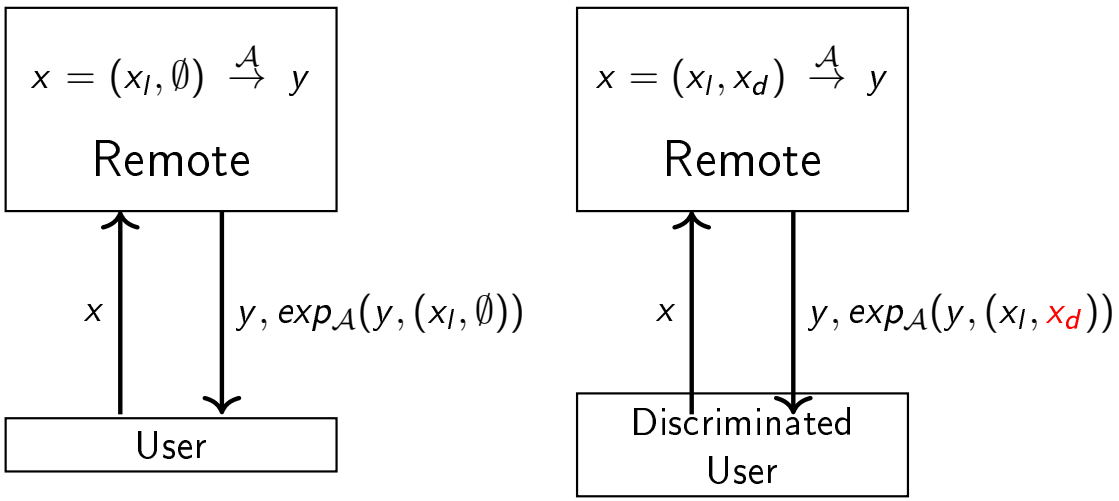


- *Black-box* classifier: provide request x, obtain decision y

Remote explainability faces the bouncer problem, Le Merrer, Tredan, Nature Mach. Int. 2020 | also cf Fairwashing

Manipulation of the platform 2/2: the bouncer problem

Audit scheme:
`auditor` finds discriminations in `platform` explanation of decisions



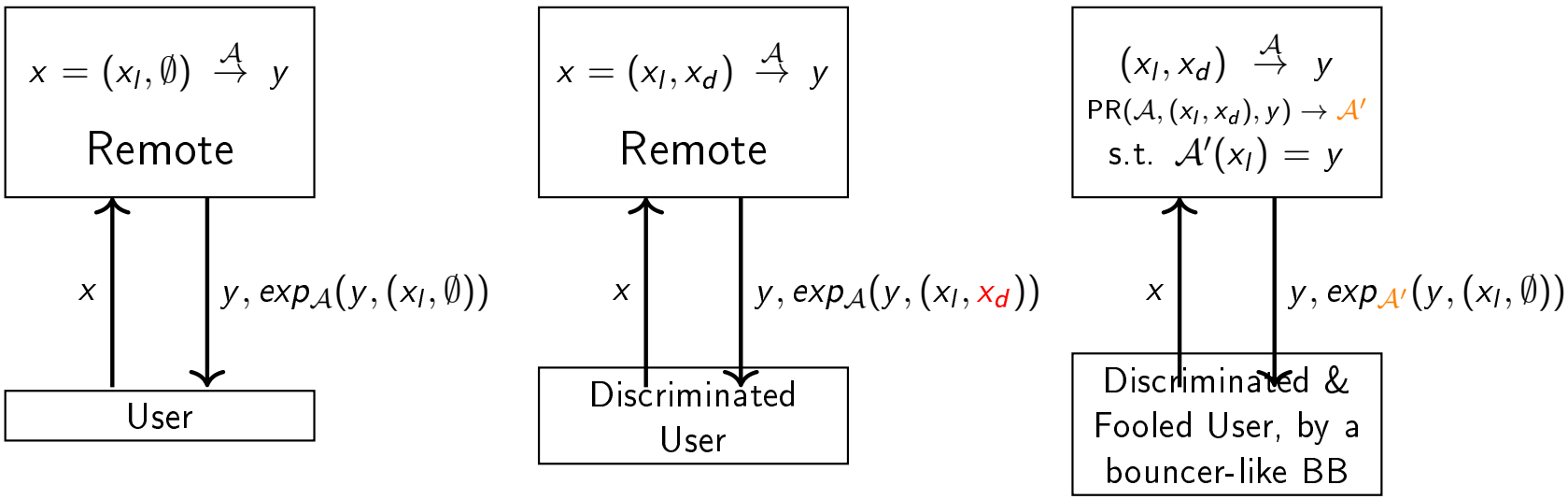
- *Black-box* classifier: provide request x, obtain decision y
- Intuition: if decision relies on discriminative variables, explanation will reveal it

Remote explainability faces the bouncer problem, Le Merrer, Tredan, Nature Mach. Int. 2020 | also cf Fairwashing

Manipulation of the platform 2/2: the bouncer problem

Audit scheme:

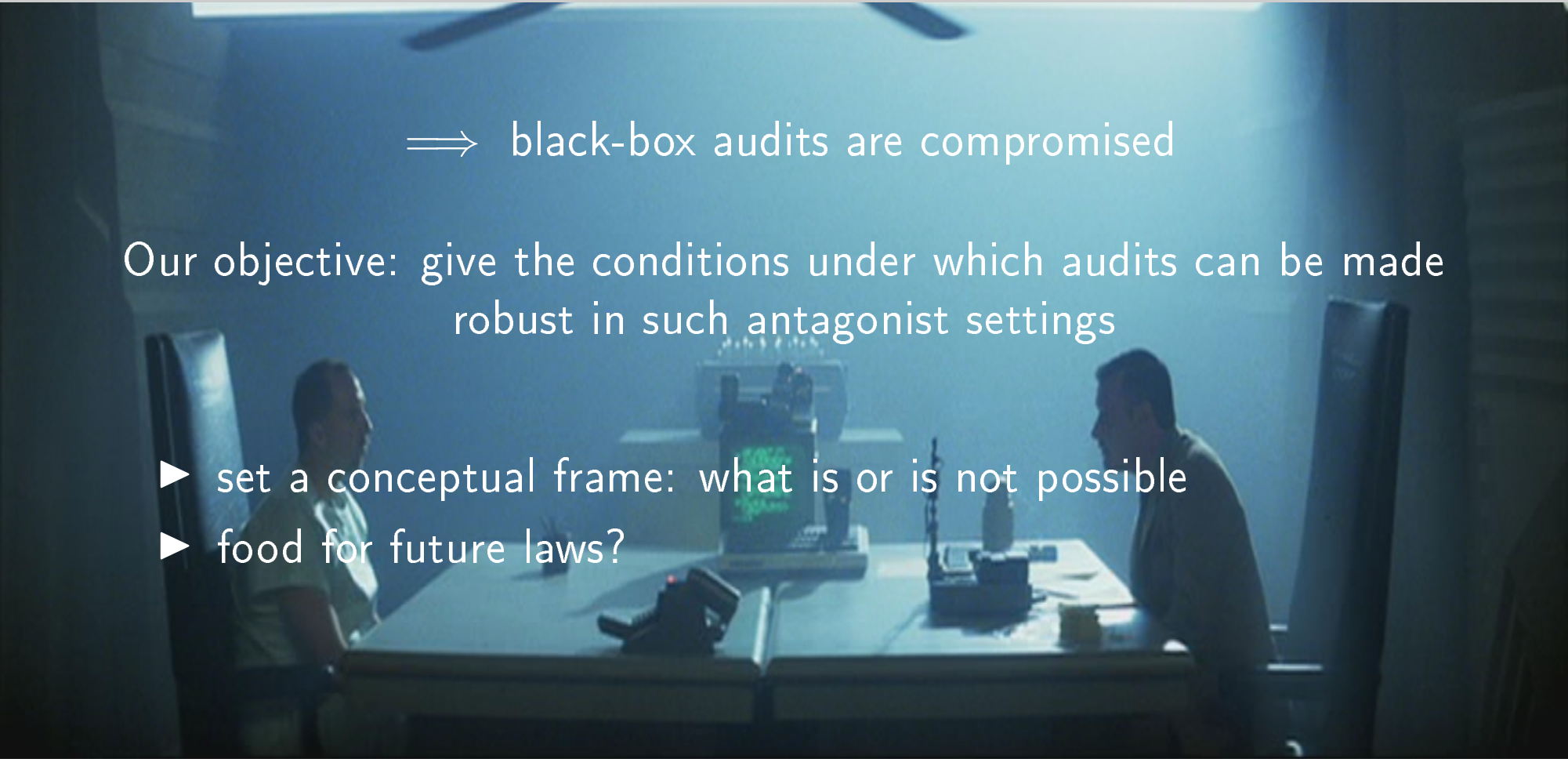
`auditor` finds discriminations in `platform` explanation of decisions



- *Black-box* classifier: provide request x, obtain decision y
- Intuition: if decision relies on discriminative variables, explanation will reveal it
- **PR attack**: generate a "legit" classifier $\mathcal{A}'$ on the fly, and explain it (like a bouncer would do...)

Remote explainability faces the bouncer problem, Le Merrer, Tredan, Nature Mach. Int. 2020 | also cf Fairwashing

Towards robust antagonistic audits

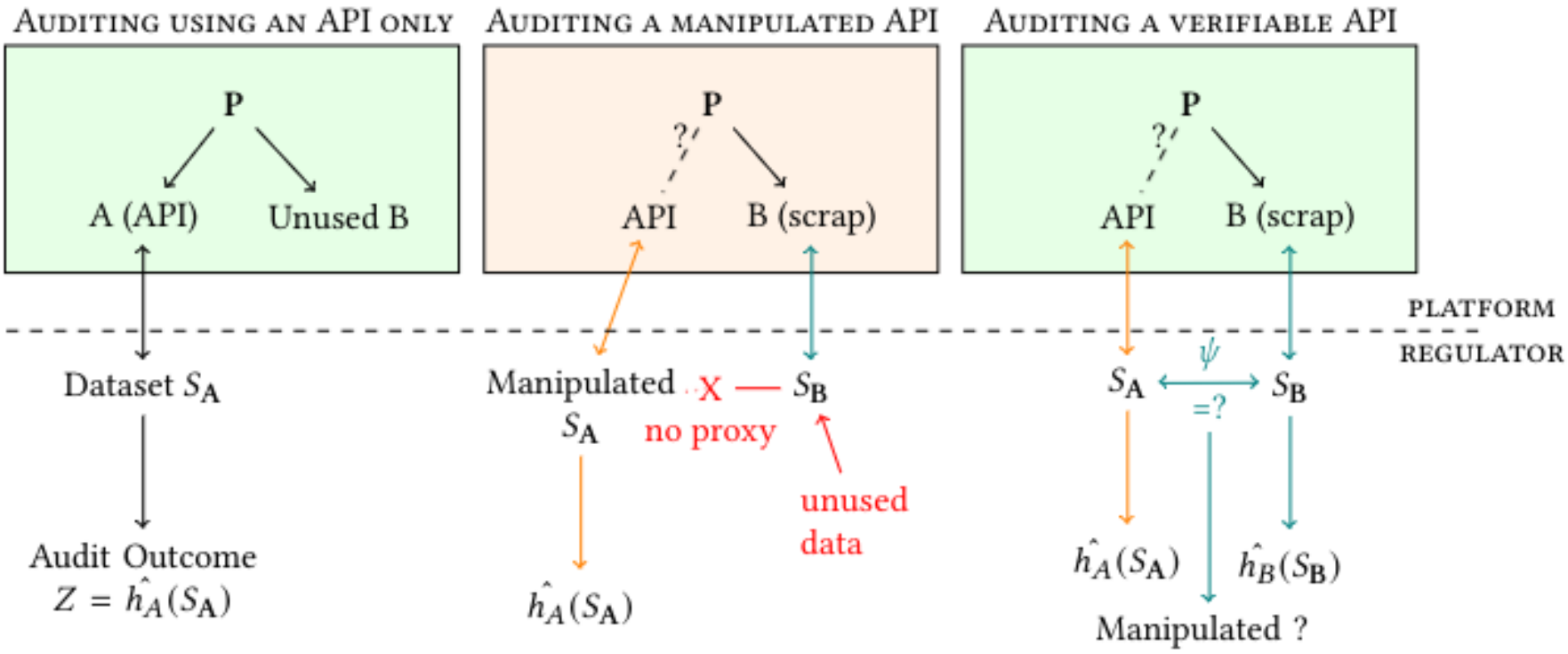


$\Longrightarrow$ black-box audits are compromised

Our objective: give the conditions under which audits can be made robust in such antagonist settings

- set a conceptual frame: what is or is not possible
- food for future laws?

Robust audits 1/2: finding inconsistencies



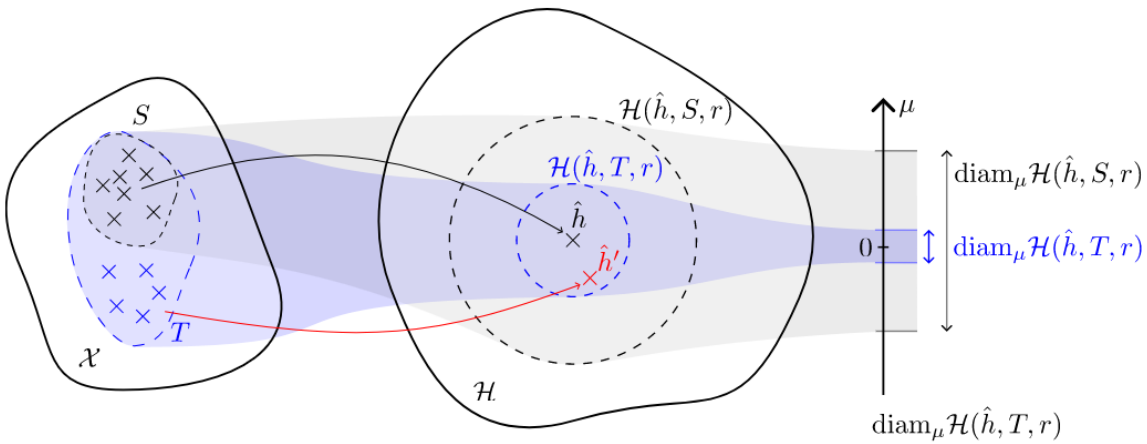
Compare observations from several sources to spot inconsistencies

Extra assumption: multiple (≥ 2) data sources

Mitigating fairwashing using Two-Source Audits, Garcia-Bourrée et al., under sub.

Robust audits 2/2: constrain the model via priors

`platform` manipulates its responses up to potential detection by the `auditor` $\rightarrow$ constrain it as much as possible



e.g. demographic parity:

$$\mu_{D_x}(\mathcal{A}) = P_{(x,x_s)\sim D_x}(\mathcal{A}(x) = 1 | x_s = 1) - P_{(x,x_s)\sim D_x}(\mathcal{A}(x) = 1 | x_s = 0)$$

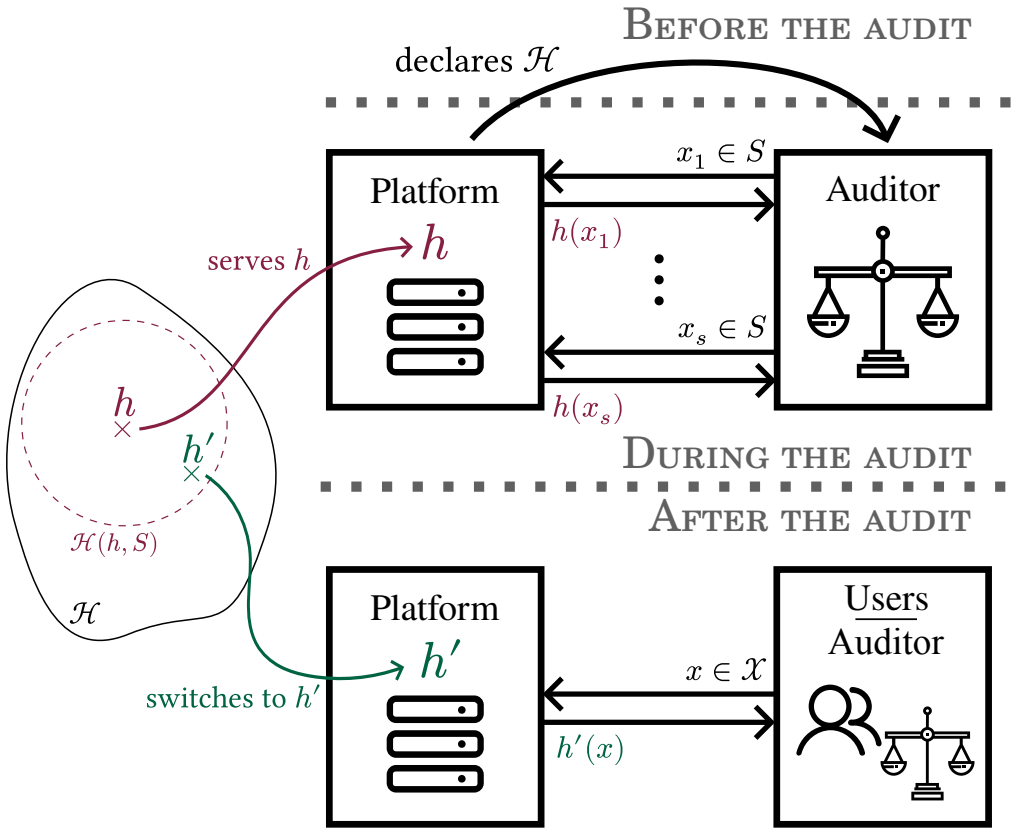
- with D_x the data distribution and x_s a sensitive attribute

Under manipulations, are some AI models harder to audit? Godot et al. SATML'24.

Robust ML Auditing using Prior Knowledge, Garcia-Bourée et al. ICML'25.

Make some assumptions: active fairness auditing, ICML'22

Constrain h to stay consistent with its previous answers

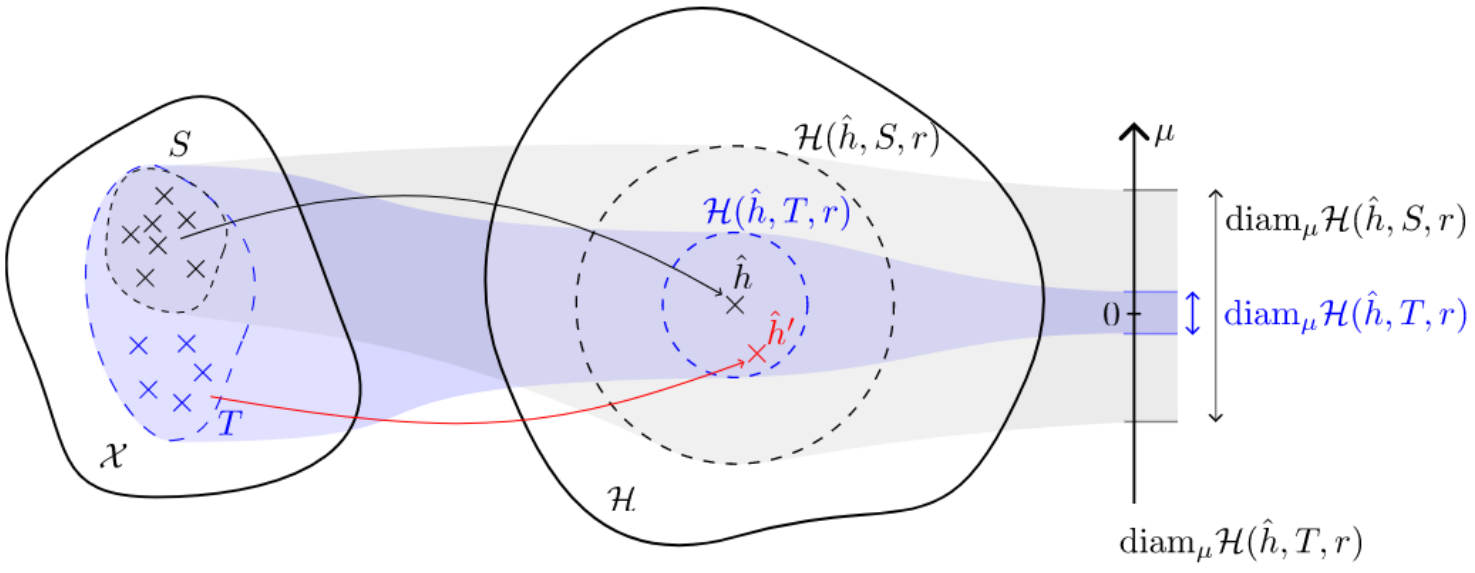


- Goal: ensure estimate within ϵ of $\mu(h_{manipulated})$
- The auditor crafts queries that constrain the model the most

Yan, Zhang, ICML 2022. Img: Augustin

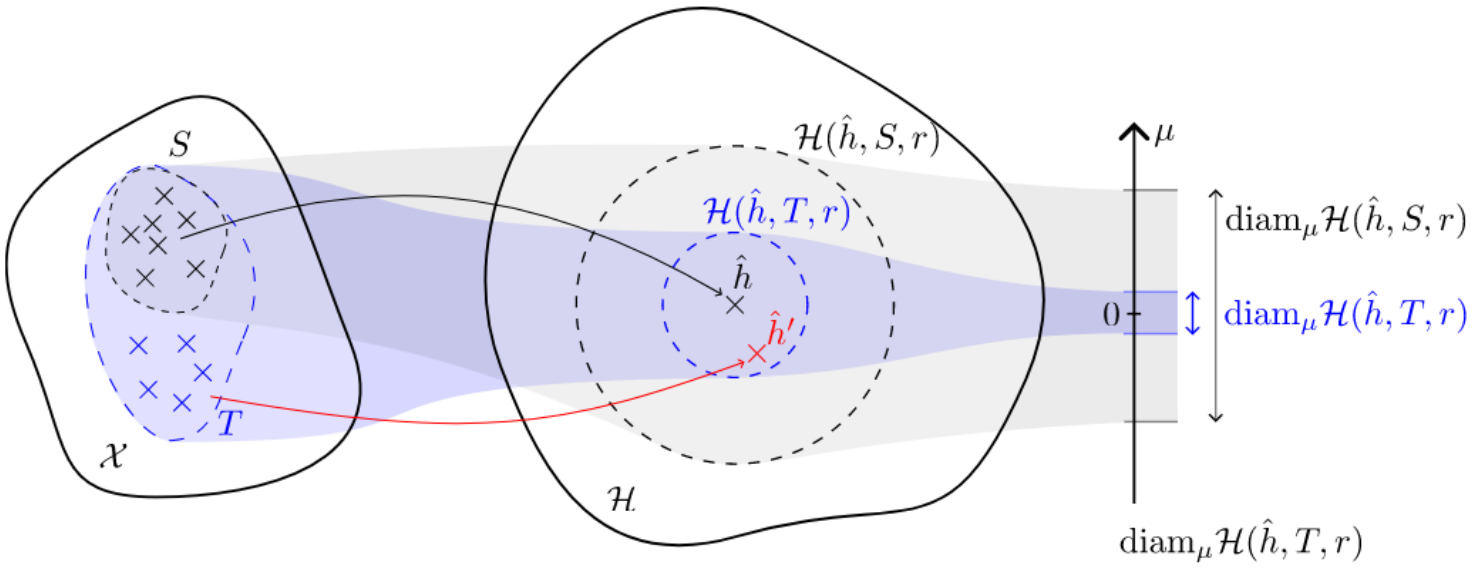
Make some assumptions: active fairness auditing

Constrain h to stay consistent with its previous answers



Make some assumptions: active fairness auditing

Constrain h to stay consistent with its previous answers



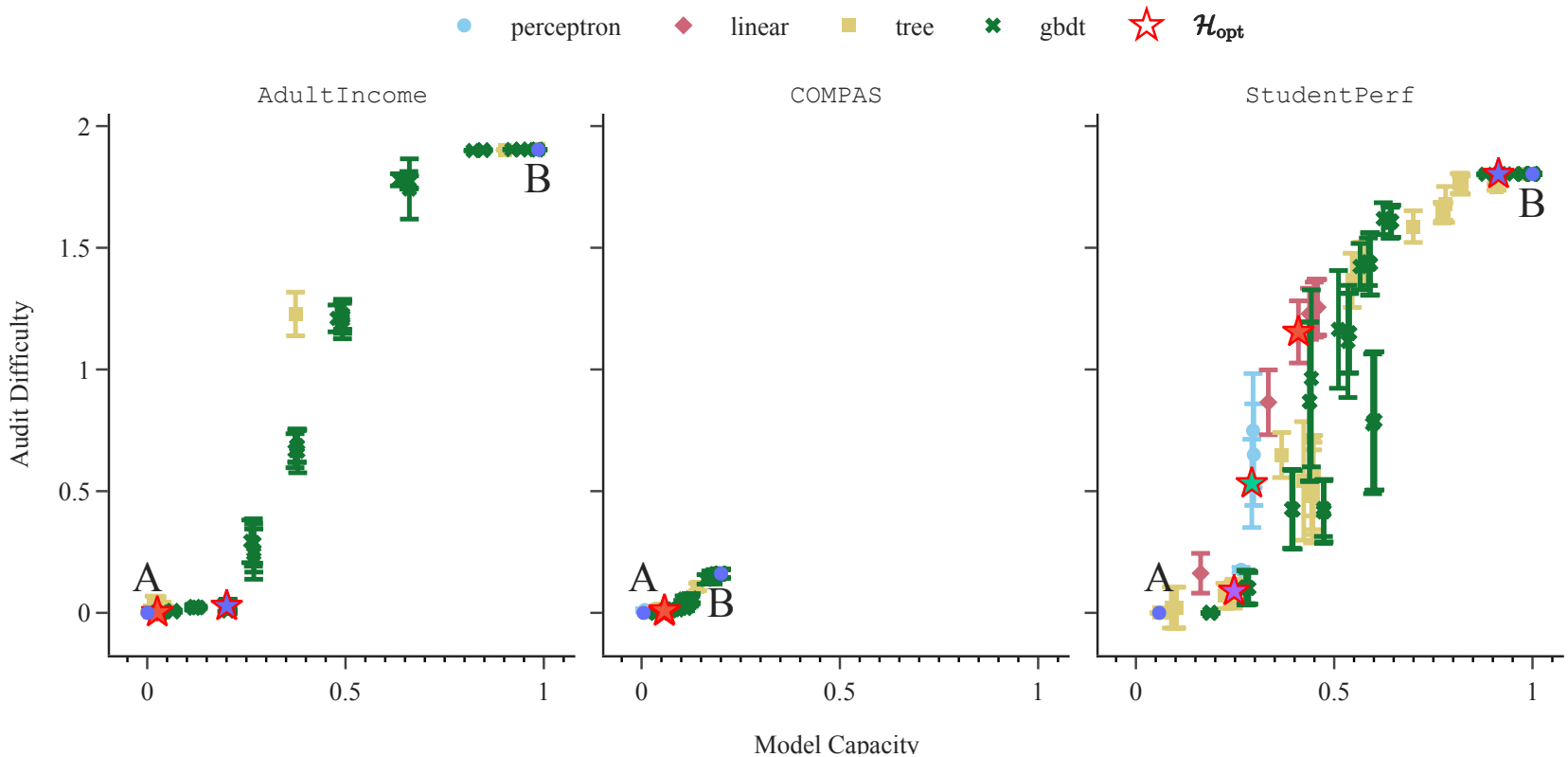
Problem: high capacity models may fit any audit set...

- Rademacher complexity as a capacity measure:
  $\mathrm{Rad}_S(\mathcal{H}) = \frac{1}{m}\mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^m \sigma_i h(z_i) \right]$, with $S = \{z_1, \ldots, z_m\}$ and σ_i random labels

Godinot et al., SATML'24.

Make some assumptions: active fairness auditing

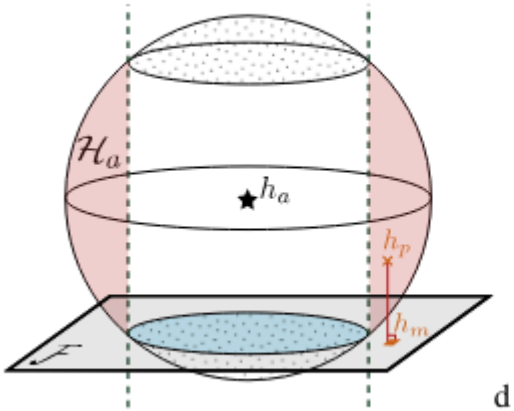
Capacity VS audit difficulty:



Current A.F.A framework not restrictive enough, regulator needs to add more constraints, ie, assumptions.

Godinot et al., SATML'24.

And one "positive" result :)



d

Figure 2. Representation of the auditor prior $\mathcal{H}_a$, the honest platform model h_p and a corresponding malicious model h_m on the fair $\mathcal{F}$ plane. The red area represents the area where platforms optimal manipulations are detected as dishonest: they fall outside of the blue region of $\mathcal{F}$

optimal manipulation is the projection of h_p on $\mathcal{F}$:

$$h_m^* = \text{proj}_{\mathcal{F}}(h_p) = \arg\min_{h \in \mathcal{F}} d(h, h_p). \quad (6)$$

The distance d in Equation (6) is the value of risk L of h using the labels of h_p as the ground truth. This scenario captures the fairwashing approach in (Aïvodji et al., 2021) in the context of explanation manipulations.

To gain intuition about the proof, we represent the audit case for $|S| = 3$ in Figure 2. By definition of the dataset prior, $\mathcal{H}_a$ is a ball of radius τ, centered on Y_a, the labels given in the audit dataset D_a. The manipulation of a model h_p can be detected only if the resulting model is outside of $\mathcal{H}_a$, as shown in orange on Figure 2. The probability of detection is thus 1 minus the volume of original models h_p whose projection on $\mathcal{F}$ lies outside on $\mathcal{H}_a$. This volume is highlighted in red in Figure 2. The detailed proof of Theorem 4.3 is deferred to Appendix A.

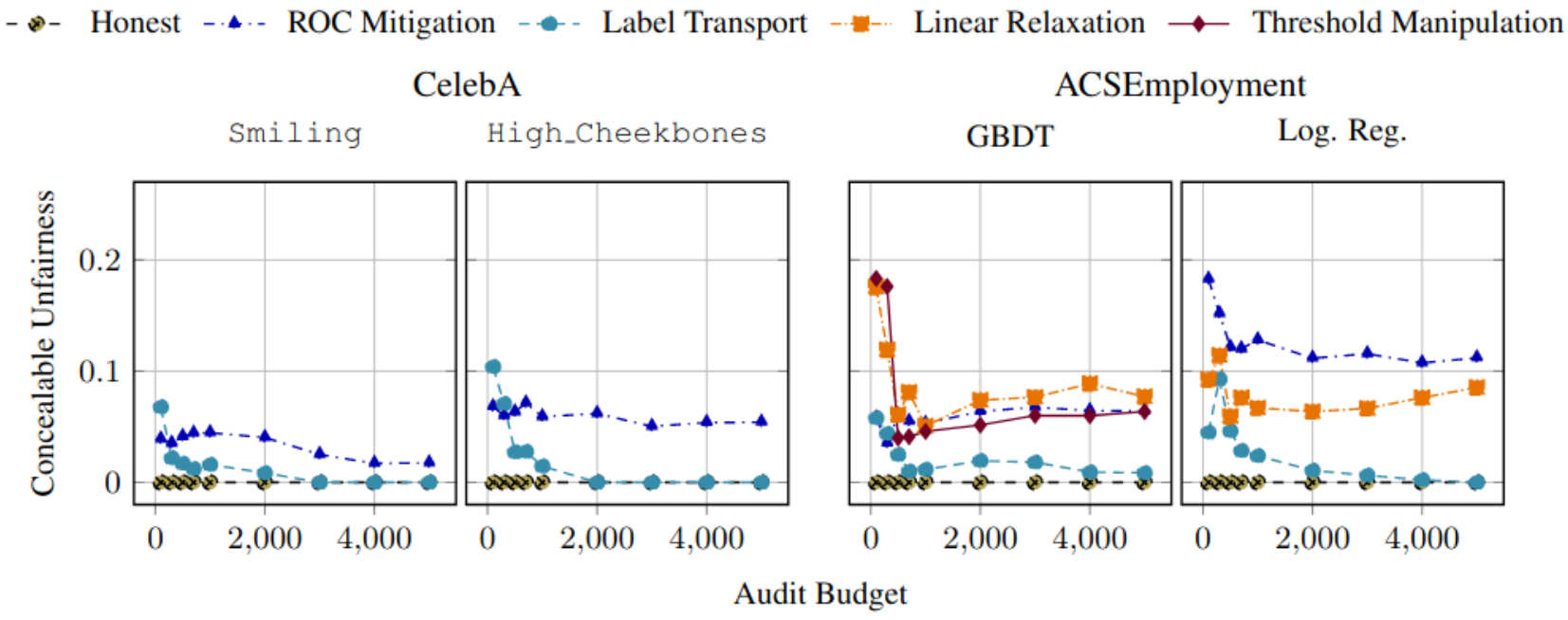
Theorem 4.3 highlights two key parameters to the auditor's success: the unfairness of the prior $\delta = d(h_a, \mathcal{F})$ and the expectability threshold τ. If the dataset prior is perfectly fair (*i.e.*, $\delta = 0$), then the auditor has no chance to detect a manipulated model as non-expectable ($P_{uf} = 0$, Corollary A.5). On the other hand, Corollary A.4 proves that, if $\tau = \delta$ [1] then $P_{uf} = 1$. Finally, in Corollary 4.4, we derive a lower bound on P_{uf} for the case $0 < \delta < \tau$. We provide the proof of Corollary 4.4 in Appendix A.

Corollary 4.4 (Detection rate lower bound)**.** *If n is even,*

$$\frac{1}{W_n}\frac{\delta}{\tau}\left(1 - \frac{\delta^2}{\tau^2}\right)^{(n-1)/2} \leq P_{uf} \leq 1.$$

Garcia Bourée et al., ICML'25.

And one “positive” result :)



Garcia Bourée et al., ICML'25.

Welcome, LLMs...

The political preferences of LLMs

David Rozado *

ECL, Otago Polytechnic, Dunedin, New Zealand

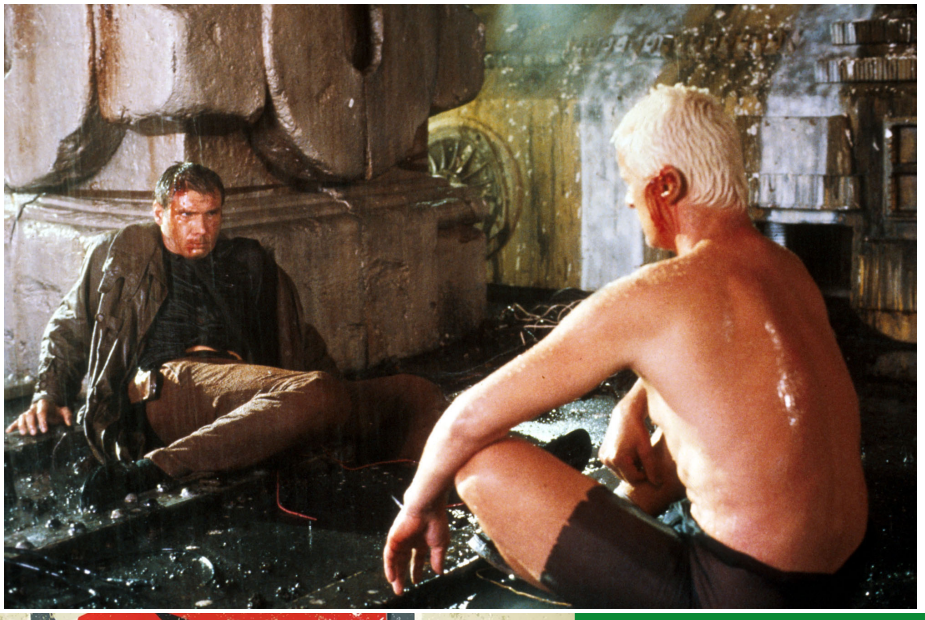
* david.rozado@op.ac.nz

Abstract

I report here a comprehensive analysis about the political preferences embedded in Large Language Models (LLMs). Namely, I administer 11 political orientation tests, designed to identify the political preferences of the test taker, to 24 state-of-the-art conversational LLMs, both closed and open source. When probed with questions/statements with political connotations, most conversational LLMs tend to generate responses that are diagnosed by most political test instruments as manifesting preferences for left-of-center viewpoints. This does

- Problems for auditors: output space is huge, they evolve fast
- Problem for all of us: *LLM-as-a-judge* paradigm, will replace search engines?, base for agentic AI ...
- Still auditable as classifiers through prompting for a decision

- Yet LLMs "capacity" is a major problem for audits. AI-2027 want AIs to watch over AIs...



- Yet LLMs "capacity" is a major problem for audits. AI-2027 want AIs to watch over AIs...

OpenAI's research on AI models deliberately lying is wild

Julie Bort · 3:54 PM PDT · September 18, 2025

- Collaboration with Pôle d'expertise de la régulation numérique (PeREN)
- enquêtes, auditions...
- Chaire SequoIA (cluster IA Rennes)

The end, in an antagonist world....

r/OpenAI • il y a 21 h
scragz

Regulating AI hastens the Antichrist, says Peter Thiel

Article

thetimes.com Ouvrir

> "because we are increasingly concerned about existential threats, the time is ripe for the Antichrist to rise to power, promising peace and safety by strangling technological progress with regulation."

2) Visibilité sur les réseaux

2) Bannissement “furtif”?

Setting the record straight on shadow banning

By Vijaya Gadde and Kayvon Beykpour

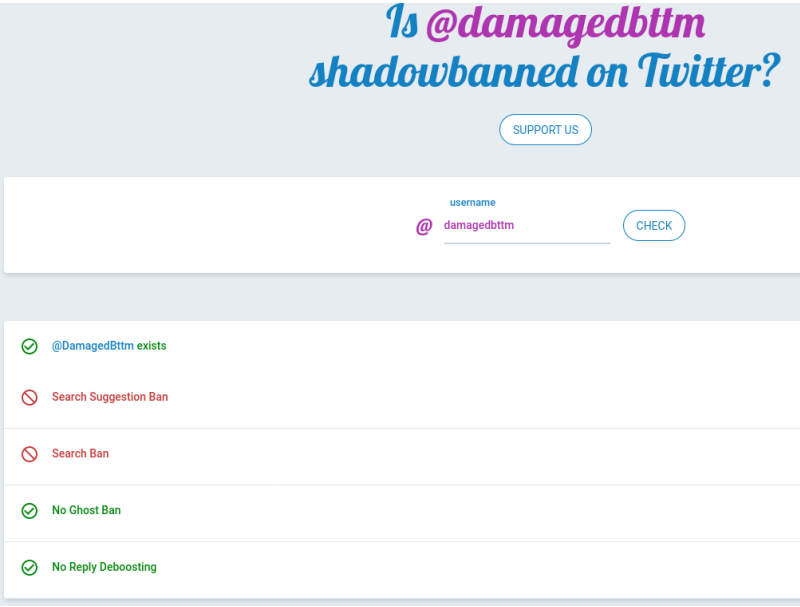
Thursday, 26 July 2018

People are asking us if we shadow ban. We do not. But let’s start with, “what is shadow banning?”

The best definition we found is this: deliberately making someone’s content undiscoverable to everyone except the person who posted it, unbeknownst to the original poster.

We do not shadow ban. You are always able to see the tweets from accounts you follow (although you may have to do more work to find them, like go directly to their profile). And we certainly don’t shadow ban based on political viewpoints or ideology.

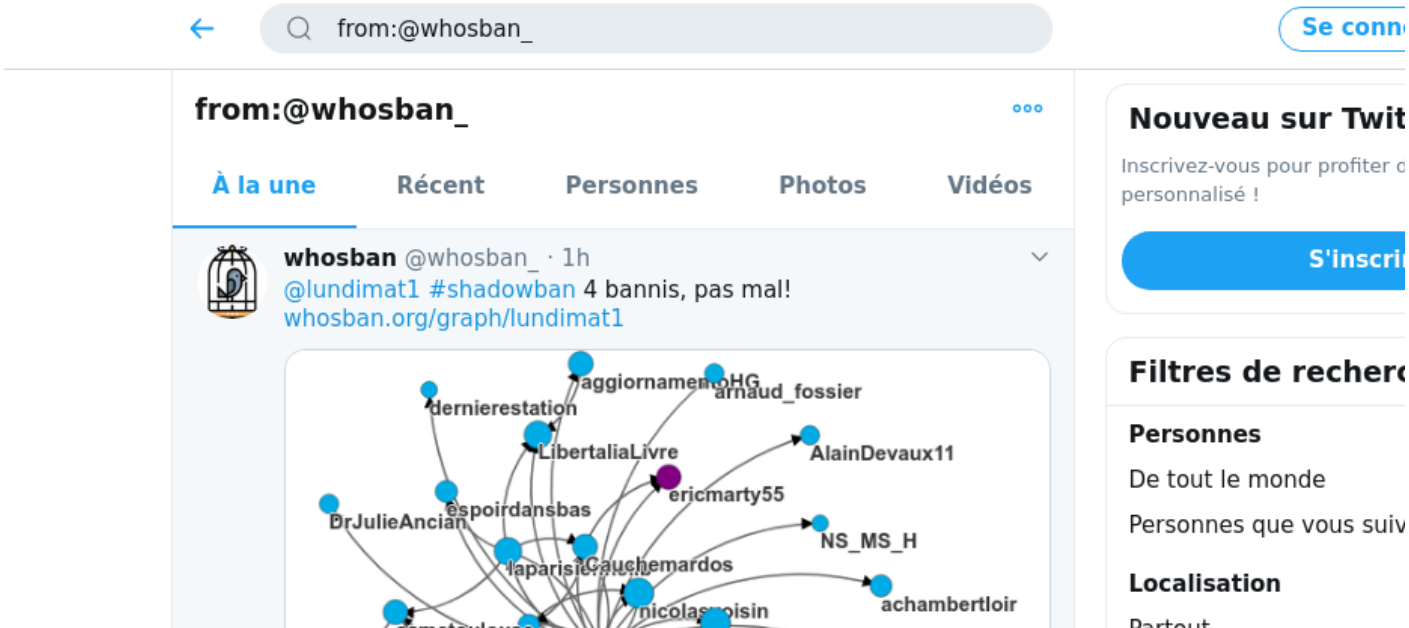
2) Collecte de preuves



Tests de visibilité par `shadowban.eu`

1. Search Ban
2. Suggestion (typeahed) Ban
3. Ghost Ban

Crawler rapide (100 profils/s)



2) Twitter: c'est un bug!

Twitter's shadow banning bug 'unfairly filtered' 600,000 accounts

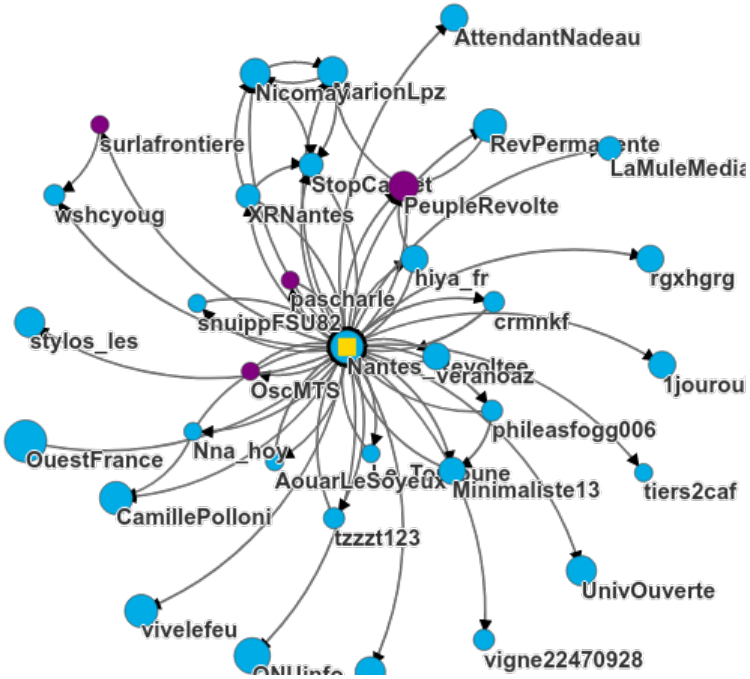
Jack Dorsey confirmed the figure to the House Energy and Commerce Committee.

Kris Holt
Contributing Reporter
Updated Wed, Sep 5, 2018 · 2 min read

2) Collecte: ego-graphes d'interaction

4 populations étudiées

1. Utilisateurs aléatoires
2. Célébrités
3. Députés
4. Robots



Ego-graphes

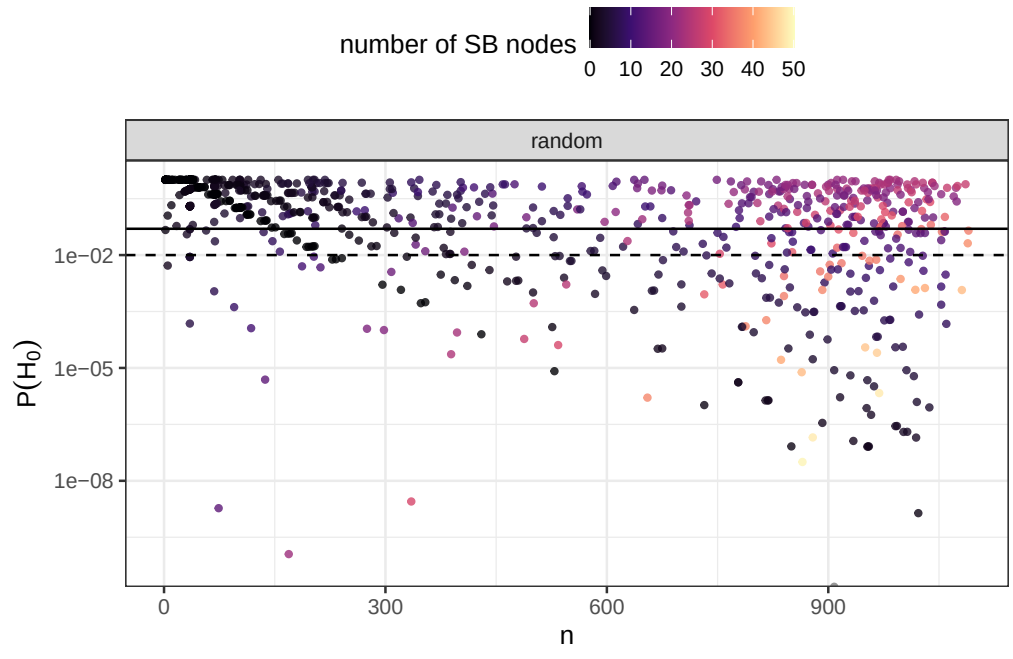
- interactions
- 33 dernières interactions, récursivement
- $\approx$ 2.5M d'utilisateurs testés

2) Prenons Twitter au mot: H_0, l'hypothèse du "bug"

Bannissement uniformément réparti

	#SB nodes	% of SB nodes/graph (avg)
FAMOUS	6,805	0.74
RANDOM	9,967	2.34
BOTS	23,358	1.97
DEPUTEES	1,746	0.50

- Plausibilité de H_0?
- Observation: $\hat{\mu} = 2.34\%$
- Modèle: urne et $|G_I|$ balles: **probabilité d'observer un tirage donné?**
- Très improbable. e.g., 'Artemis**', 703 voisins, 45.4%bannis, $P \approx 1e-315$



- Retombées: Twitter a retiré son post; question au parlement EU; journaux